Anomaly Detection and Analysis Framework for Terrestrial Observation and Prediction System (TOPS)

Petr Votava, Andrew Michaelis, Hirofumi Hashimoto
University Corporation at Monterey Bay/NASA Ames
Seaside, CA, United States
{pvotava, armichae, hhashimo}@nasa.gov

Ramakrishna Nemani
NASA Ames Research Center
Moffett Field, CA, United States
rnemani@nasa.gov

Abstract— **Terrestrial Observation and Prediction System (TOPS) is a flexible modeling software system that integrates ecosystem models with frequent satellite and surface weather observations to produce ecosystem nowcasts (assessments of current conditions) and forecasts useful in natural resources management, public health and disaster management. We have been extending the Terrestrial Observation and Prediction System (TOPS) to include capability for automated anomaly detection and analysis of both on-line (streaming) and off-line data. While there are large numbers of anomaly detection algorithms for multivariate datasets, we are extending this capability beyond the anomaly detection itself and towards an automated analysis that would discover the possible causes of the anomalies. There are often indirect connections between datasets that manifest themselves during occurrence of external events and rather than searching exhaustively throughout all the datasets, our goal is to capture this knowledge and provide it to the system during automated analysis, which results in more efficient processing. Finally, the project goal is to provide a testbed for a number of anomaly detection and data-mining algorithms and packages, so that they can be tested on large volumes of spatio-temporal datasets.**

Keywords- TOPS; RDF; data-mining; anomaly detection; OWL; MODIS; Landsat; NASA

I. INTRODUCTION

The management and processing of Earth science data has been gaining importance over the last decade due to higher data volumes generated by a larger number of instruments, and due to the increase in complexity of Earth science models that use this data. The volume of data itself is often a limiting factor in obtaining the information needed by the scientists; without more sophisticated data volume reduction technologies, possible key information may not be discovered. We are especially interested in automatic identification of disturbances within the ecosystems (e,g, wildfires, droughts, floods, insect/pest damage, wind damage, logging), and focusing our analysis efforts on the identified areas. There are dozens of variables that define the health of our ecosystem and both long-term and short-term changes in these variables can serve as early indicators of natural disasters and shifts in climate and ecosystem health. These changes can have profound socio-economic impacts and we need to develop capabilities for identification, analysis and response to these changes in a timely manner. Because the ecosystem consists of a large number of variables, there can be a disturbance that is only apparent when we examine relationships among multiple variables despite the fact that none of them is by itself alarming. We have to be able to extract information from multiple sensors and observations and discover these underlying relationships. As the data volumes increase, there is also potential for large number of anomalies to "flood" the system, so we need to provide ability to automatically select the most likely ones and the most important ones and the ability to analyze the anomaly with minimal involvement of scientists. We describe a prototype architecture for anomaly driven data reduction for both near-real-time and archived products in the context of the Terrestrial Observation and Prediction System (TOPS)[1], a flexible modeling software system that integrates ecosystem models with frequent satellite and surface weather observations to produce ecosystem nowcasts (assessments of current conditions) and forecasts useful in a range of applications including natural resources management, public health and disaster management.

II. BACKGROUND

The Terrestrial Observation and Prediction System (TOPS) is a flexible modeling software system that integrates ecosystem models with frequent satellite and surface weather observations to produce ecosystem nowcasts (assessments of current conditions) and forecasts useful in a range of applications including natural resources management, public health and disaster management. TOPS integrates and pre-processes NASA’s Earth Observing System (EOS) data so that land surface models can be run in near-real-time (as well as retrospectively) with minimal user intervention, providing for accurate and timely interpretation of EOS data. We are applying nowcasts and forecasts from TOPS to assist with identification of vulnerabilities of different socio-economic and resource management approaches to fluctuations within our biosphere, and assessing mitigation for potential negative impacts. We are currently working with the National Park Service (NPS), U.S. Geological Survey, Department of Water Resources of California, SERVIR (the Regional Monitoring and Visualization System for Mesoamerica), the Center for Disease Control (CDC), the Environmental Protection Agency (EPA), commercial firms including MDA Federal

Inc., and Robert Mondavi wineries. We produce about 30 different products on daily basis ranging from satellite-based land surface properties, gridded weather fields to modeled ecosystem fluxes. Over the past several years TOPS has provided over 20 TB of data to our application partners and the Earth science community. The TOPS data archive is currently around 300TB.

III. System Overview

The anomaly detection component of the TOPS system is designed as a plug-in framework, because apart from our own implementations of anomaly detection algorithms, our goal is to establish a test bed for anomaly detection into which 3rd party algorithms can be integrated with ease. The framework is an integration point that provides access to databases containing large volumes of satellite, climate and model data and number of pre-processing tools for converting data to formats required for a number of different data-mining and anomaly detection packages. The framework will also interact with the anomaly database, which contains results of previous data analysis, data models generated from training data and other data statistics useful for future analysis. Additionally, we are currently generating global climatology datasets from MODIS[2] level 3 and 4 land products at native resolution, which will be available as a baseline for the anomaly detection framework. The integration of the anomaly detection framework with TOPS is depicted in Figure 1.

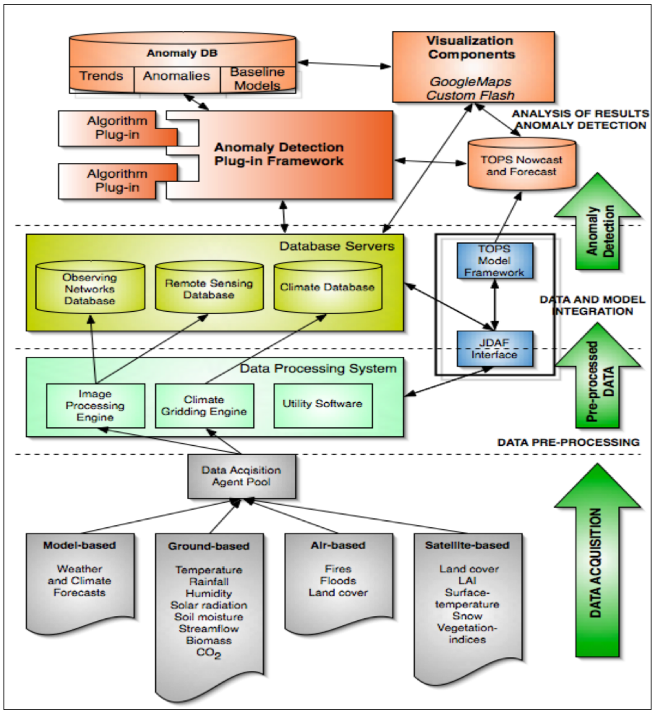

Figure 1 - Anomaly Framework integration with TOPS

The key components of the anomaly detection framework consist of *Knowledge Base* for description of TOPS data and model relationships; *Control Module* for improved automation of repeated analysis and model execution; *Anomaly Detection Framework*, which provides the infrastructure for easy extensibility with 3rd party anomaly detection software; *Anomaly Confirmation Component* that implements automatic anomaly confirmation in similar datasets and models; *Anomaly Database*, and finally the *Anomaly Analysis Component* that performs a pre-defined automatic analysis triggered by the anomaly detection. The overview of the framework is depicted in Figure 2.

A. *Knowledge Base*

The *Knowledge Base* captures notions of data and model relationships, such as similarity, and compatibility, so that other components can automatically select similar datasets during anomaly analysis or confirmation, and have the ability to use different regional models over different areas. Apart from containing data and model descriptions the *Knowledge Base* also provides query capabilities, so that similar data or models can be readily identified. We are populating the *Knowledge Base* system with data hierarchy for 30+ different datasets that are already used within TOPS. TOPS models will be described in a similar fashion where their inputs and outputs will be referencing datasets in the data hierarchy. Additional requirement on the model descriptions will be regional constraints that will make it easier to automatically select different models for different regions as we move towards analysis of specific anomaly within a smaller geographic area and we want to execute more specialized model. The data hierarchy is described using Resource Description Framework (RDF) [3] markup language. The descriptions are stored using Sesame[4] RDF server as an repository infrastructure, and an implementation of SPARQL[5] query engine to provide the desired query capabilities.

B. *Control Module*

The *Control Module* implements the logical flow of the automation process and it is built on top of existing TOPS architecture. First, it interfaces existing TOPS data acquisition system and provide capabilities for a selective dynamic data acquisition that can be performed on-demand. Second, given temporal, spatial and dataset constraints, it can form queries that will be targeted for the *Knowledge Base* to retrieve compatible model/data combinations. Finally, the *Control Module* can initiate the execution of the model/data combinations retrieved in the previous query. The execution can be done in parallel because each data/model combination is independent.

C. *Anomaly Detection Framework*

The *Anomaly Detection Framework* provides infrastructure for deploying algorithms for both on-line (streaming) and off-line (archived) anomaly detection in multi-spectral multi-variable spatio-temporal datasets from the climate change and ecosystem domains. We are testing this framework with existing implementations for both on-line and off-line data processing based on algorithms developed at NASA Ames IVHM Data-mining lab. Our goal is the adaptation of these algorithms to the domain of Earth science, particularly addition of a spatial component for analysis of gridded data, such as satellite images. These algorithms are capable of performing on-line anomaly detection on gigabyte size data streams, such as near-real time data from the EOS direct broadcast stations that are currently processed by TOPS, as well as on multi terabyte datasets archived by TOPS. We aim to have the ability to integrate number of 3rd party anomaly-detection and data-mining software such as Weka[6] or Apache Mahout[7] into this framework. We hope that this can become a testbed for a faster development and testing of new algorithms developed by the data-mining community at NASA and elsewhere aimed at fast anomaly detection in very large Earth science datasets.

D. Anomaly Confirmation Component

There are many occasions, when detected anomaly is not related to ecosystem health, but rather noise in the data, faulty sensors, cloud contamination, and other factors, and as there is a cost associated with analysis of the discovered anomalies, we want to be confident (with certain probability) that we are focusing on the correct ones. This will be accomplished by the *Anomaly Confirmation Component* that will be able us to re-run anomaly detection algorithm on similar datasets as defined in the knowledge base to confirm the occurrence of the anomaly. In case the anomaly was discovered in an output of one of TOPS models, we will have two courses of action – first, we will re-run the model with different data sources and check for the presence or absence of the anomalies in the results, and second, we will check for anomaly in the inputs to the model, which has the potential of giving us a hint of what caused the anomaly in the first place.

E. Anomaly Analysis Component

Once the anomaly is identified and confirmed, we will classify it using our existing database and determine a set of possible actions. The actions will range from new data acquisition and analysis using standard image processing and statistical methods, to initiation of a new higher-resolution regional model runs. The execution of the response will be responsibility of the *Anomaly Analysis Component*, which provides the initiation and the coordination of the analysis plans. This component will utilize a repository of pre-defined workflows based on past analysis and encoded using VisTrails[8] workflow engine.

IV. OPERATIONAL CONCEPT

The system can be used for two separate purposes. First, it can be used for anomaly detection in on-line (streaming data) in near-real-time. And secondly, it can be used for off-line anomaly detection in large data archives to identify disturbances in time-series multidimensional datasets.

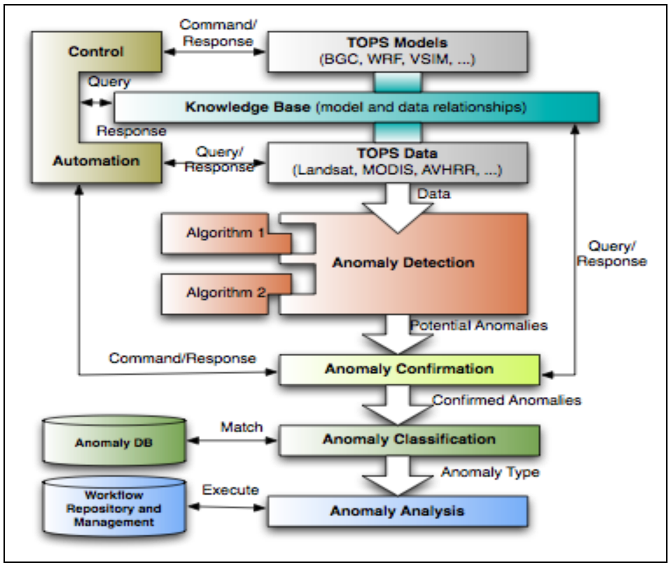

Figure 2 - Overview of the Anomaly Detection Framework

While the initiation of this process differs, the rest of the operations remain the same. The data are passed through the *Anomaly Detection Framework* onto one of the detection algorithm implementations, which may identify number of anomalies in the dataset that will typically consist of number of different variables that are co-located in both time and space. The user can specify number of parameters that will limit the results, or provide specific distance metric to be used for neighbor-related anomaly detection algorithms. Once the anomalies are returned, each anomaly is verified using the *Anomaly Confirmation Component*. During this process the *Knowledge Base* is queried for definition of similar variables to those that caused the anomaly to be detected. If the variables are not locally available, the *Control Module* will initiate data acquisition. Once the data are available, they will be combined in the same way as the original data and passed back to the *Anomaly Detection Component*. The user can specify how long should this process be repeated and what are the boundaries for proceeding further with analysis. There are other possibilities driven by the *Anomaly Confirmation Component* – if the anomaly is detected in model output, different inputs to the model are obtained and the model is re-run using the *Control Module*, and the model output is tested for anomalies. Finally, if the anomaly is present in the model output, the inputs into the model can be combined and passed to the *Anomaly Detection Component,* which can help us to determine the original cause of the anomaly. Once the confirmation process is concluded, there are two possibilities. If anomaly is not confirmed, the information will be logged into a database to keep track of the detections for possible future re-analysis. If the anomaly is verified, the information related to the anomaly is used to query the *Anomaly Database* in order to determine the type of the anomaly and proper response plan. If there is no approximate match in the *Anomaly Database*, the user is notified and at this point he or she can choose to analyze the anomaly and determine course of action if same anomaly is identified in the future. If the anomaly matches, we can use the reference returned by the *Anomaly Database* to dispatch appropriate action based on workflow(s) defined in response to this particular anomaly type.

V. Status

We have completed the ontology definition and the *Anomaly Detection Framework* and integrated a number of TOPS existing anomaly detection algorithms into the system. We are currently testing it against TOPS operational satellite, climate and model output datasets. We are currently also completing work on the *Anomaly Confirmation* component Finally, we are preparing a workflow repository with sample workflows that will be triggered in response to detected anomalies. The system should be completed in early 2012.

The framework is written in combination of Java and Python, while the anomaly algorithm and other software packages are integrated using thin wrappers.

Acknowledgment

This work has been funded under NASA's Advanced Information System Technologies (AIST) Program under the NNH08ZDA001N –AIST NRA.

References

[1] R.R. Nemani, et al., "Terrestrial Observation and Prediction System (TOPS): Developing ecological nowcasts and forecasts by integrating surface, satellite and climate data with simulation models," : Research and Economic Applications of Remote Sensing Data Products, Eds : U. Aswathanarayana, Taylor & Francis Book Series, London, 2007.

[2] C.O. Justice, et al.. "The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research," *IEEE Transactions on Geoscience and Remote Sensing*, 36(4), 1998, pp.1228-1249.

[3] S. Powers 2003, "Practical RDF," O'Reilly Media, Inc. 2003.

[4] J. Broekstra, A. Kampman, F. van Harmelen, "Sesame: A Generic Architecture for Storing and Querying RDF," International Semantic Web Conference 2002, Sardinia, Italy.

[5] E. Prud'hommeaux, A. Seaborne, "SPARQL Query Language for RDF : W3C Recommendation," 2008.

[6] I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques (Third Editions)," Morgan Kaufmann, January 2011.

[7] S. Owen, R. Anil, T. Dunning, E. Friedman, "Mahout in Action," Manning Publishing, June 2010.

[8] S.P. Callahan, J. Freire, E. Santos, C.E. Scheidegger, C.T. Silva, H.T. Vo, "VisTrails: Visualization meets Data Management," Proceedings of ACM SIGMOD, 2006.